SPECIFICATION

Electronic Version 1.2.8
Stylesheet Version 1.0

SYSTEMS AND METHODS FOR TRIAGE OF PASSAGES OF TEXT OUTPUT FROM AN OCR SYSTEM

Background of the Invention

[Field of Invention]

[0001] This invention generally relates to processing text passages that are subjected to character recognition processes.

[Description of Related Art]

[0002] Digitizing paper documents generally involves creating a bitmap image of the paper document using a scanner or similar device and then storing the bitmap image in a computer system. To retrieve and evaluate bitmap images, the computer must recognize characters within the bitmap image created by the scanner. Character recognition techniques, for example, optical character recognition (OCR) techniques, are generally used to convert images of characters, usually provided to the computer system in some standard format, such as, for example, the tagged image file format (TIFF), into machine-legible coded form of those characters, such as, for example, ASCII or Unicode.

[0003] In carrying out this conventional conversion process, some fraction of the characters may not be converted correctly. Some basic OCR errors include, for example: substitution, where one character in a text passage is mistaken for another; deletion, where the correct character is missing; and insertion, where a spurious character is introduced. Often times, post-OCR correction of the document image must be performed in order to maintain acceptable document content accuracy.

Summary of the Invention

[0004] The cost of conversion of paper documents, through image acquisition, for example, scanning, and recognition, for example, OCR, into machine-legible coded form, such as, for example, ASCII or Unicode, is often dominated by the expense of post-OCR correction. At times, post-OCR correction, which may have to be performed manually, can require rekeying of the entire document, thus obviating every advantage of OCR. This occurs because the present state of the art of OCR can only rarely yield uniformly high accuracies across collections of dissimilar documents, for example, those containing a variety of typefaces, languages, layout formats, and/or image qualities.

[0005] This invention provides systems and methods for automatic triage of text passages outputted from a character recognition system, for example, an OCR system, using trainable models of the accuracy of the system that are based on attributes of the individual characters of each text passage. To triage is to decide what resources will be used to improve, for example heal or fix, something. In various exemplary embodiments according to this invention, triage means the decision procedure for making a triage decision, and the effects of that decision on later processing. In this invention, the item to be improved, on which triage decisions are made, are the results of running optical recognition processes (OCR) on images of documents containing passages of text. Thus, triage is the process of rapidly and automatically estimating the quality of OCR to enable procedural decisions later in the scanning and recognition pipeline or process.

[0006] This invention provides systems and methods for automatic triage of OCR-outputted text passages to determine the best post-OCR processing step required for the text passages evaluated. In various exemplary embodiments, post-OCR processing steps may include, for example, sending the OCR-output text passage directly to the end user without any post-OCR processing, such as, for example, rekeying and/or correction of text passage. In alternate exemplary embodiments, post-OCR processing steps may include, for example, sending the OCR-output text passage through a post-OCR inspection and correction stage. In various alternate exemplary embodiments, post-OCR processing steps may include, sending the original text passage image to be completely keyed in manually. In various alternate

exemplary embodiments, post-OCR processing steps may include a combination of inspection, correction and manual rekeying steps.

[0007] This invention provides systems and methods for applying OCR text triage tools fully automatically and at high speed in a "scan-and-convert" document service center setting to improve accuracy, speed throughput, raise productivity, improve quality assurance, lower production costs, and provide management with a real-time decision tool for selecting best operating practices.

[0008] In various exemplary embodiments, the systems and methods according to this invention determine the accuracy of an OCR-outputted document text passage.

[0009] In various exemplary embodiments, the systems and methods according to this invention automatically triage an OCR-output text passage by determining at least one OCR-output character attribute for each character within the OCR-output text passage, determining an error rate for the OCR-output text passage using a triage model and the determined at least one OCR-output character attribute, and comparing the determined error rate for the OCR-output text passage with an OCR-output text passage threshold error rate to perform an OCR-output text passage triage decision.

[0010] In various exemplary embodiments, the systems and methods according to this invention automatically triage OCR-output text passages that include at least one of pages, characters, words, phrases, text-lines, sentences, paragraphs, columns of text, blocks of text, text articles, multi-page documents, collections of single-page documents, collections of multi-page documents, and the like.

Brief Description of the Drawings

[0011] Various exemplary embodiments of the systems and methods of this invention will be described in detail below, with reference to the following figures, in which:

[0012] Fig. 1 illustrates a high level document scan-and-convert network environment;

[0013] Fig. 2 is a functional block diagram of one exemplary embodiment of a system for automatic triage of text passages outputted from an OCR system according to this invention;

[0014] Fig. 3 provides examples of the types of OCR-output text passages that can be evaluated according to this invention;

[0015] Fig. 4 provides examples of the types of OCR character attributes that can be evaluated according to this invention;

[0016] Fig. 5 is a functional block diagram showing in greater detail one exemplary embodiment of the trained off-line triage model of Fig. 2, according to this invention;

[0017] Fig. 6 is a flowchart outlining one exemplary embodiment of a method for training an off-line model usable to determine statistical models of class conditional character attribute according to this invention;

[0018] Fig. 7 is a high-level schematic representation of one exemplary embodiment of the implementation of the method for triage of OCR-output text passages according to this invention;

[0019] Fig. 8 is a flowchart outlining one exemplary embodiment of a method for on-line triage of an OCR-output text passage according to this invention;

[0020] Fig. 9 is a flowchart outlining in more detail one exemplary embodiment of one of the steps of the method for triage of an OCR-output text passage of Fig. 8; and

[0021] Fig. 10 provides examples of the types of OCR-output text characters that can be evaluated according to this invention.

Detailed Description of the Methods

[0022] Converting paper documents into a machine-legible coded format produces OCR-outputted documents having various degrees of accuracy across collections of dissimilar documents. Existing systems and methods for identifying the accuracy of OCR-outputted documents on the basis of, for example, confidence values for text characters recognized, or spell-checking a document, are unsophisticated, time consuming and labor intensive.

[0023] Automatically deciding which of the documents emerging from the OCR process possess acceptable accuracy could allow these documents to skip manual correction, lowering production costs. Furthermore, automatically identifying OCR-outputted

documents having extremely low accuracy rate values would allow the documents to be completely rekeyed, which could also lower production costs since it eliminates the tedious step associated with manually verifying the OCR-outputted text. We call this a 'triage' decision, in analogy with the practice of emergency medical rescue staff who are trained to classify injured individuals into those who need immediate medical attention and those who do not. Medical triage tries to maximize the number of survivors given finite care facilities. OCR triage tries to maximize the number of pages that skip correction given a fixed uniform accuracy target.

[0024] The systems and methods of this invention enable scan-and-convert businesses and alike to determine the accuracy of an OCR-outputted document text passage and then to automatically decide the best post-OCR processing step required for that particular text passage. Post-OCR processing steps may include, for example, sending the OCR text passage directly to the end user without any post-OCR rekeying or correction, sending the OCR text passage through a post-OCR inspection and correction stage, sending the original text passage image to be completely keyed in manually, or a combination thereof.

[0025] Fig. 1 shows one exemplary embodiment of a network environment 100 that the systems and methods of this invention are usable with. As shown in Fig. 1, in the scan-and-convert network environment 110, hard-copy documents, such as, for example, document pages 120, 122, 124 and 126, are scanned into one or more document scanning units, for example, optical scanning units 130, 132. Hard-copy documents 120, 122, 124 and 126 typically include images representing alphanumeric characters. The optical scanning units 130 and 132 scan the hard-copy documents 120, 122, 124 and 126 and produce electronic signals corresponding to a digital page image representing the image of hard-copy documents 120, 122, 124 and 126. One or more scanning / processing / storage stations 140 and 142, such as, for example OCR scanning / processing / storage stations 140 and 142, are provided to receive the digital page image, perform any necessary or desired character recognition functions, and store the OCR-outputted documents or text passages 150. The OCR scanning / processing / storage stations 140 and 142 may be connected to the optical scanning units 130 and 132 via links 135 and 137, respectively, or through other connections present within network 110.

[0026] A user, using a personal computer or other device that is equipped with a suitable communications software, can access the network 110 over a communication link 214 and is able to access the OCR-outputted documents or text passages 150 available on the network 110. The network 110 includes, but is not limited to, for example, local area networks, wide area networks, storage area networks, intranets, extranets, the Internet, or any other type of distributed network, each of which can include wired and/or wireless portions.

[0027] The large volume of OCR-outputted documents or text passages 150 available on the scan-and-convert network 110 presents significant difficulties to a user in manually determining which of the text passages 150 meet the text passage error threshold values imposed by the customer for the group of scanned documents. In various exemplary embodiments, a network or web-connected OCR output text triage system 200 according to this invention allows the text passages 150 to be automatically triaged in an accurate and efficient manner such that the number of pages or text passages that skip correction, given a fixed uniform accuracy target, is maximized. The text passages 150, as shown in Fig. 3, to which the systems and methods of this invention are applied include, for example, entire pages, individual text characters contained within a page, words, phrases, text-lines, sentences, paragraphs, columns of text, blocks of text, text articles, multi-page documents, collections of single-page documents, collections of multi-page documents, and the like.

[0028] As shown in Fig. 1, in various exemplary embodiments, the OCR output text triage system 200 may receive the OCR-outputted document text passages from the scanning / processing / storage stations 140 and 142 via the scan-and-convert network environment 110. Alternately, the OCR output text triage system 200 may be connected directly to the scanning / processing / storage stations 140 and 142. Moreover, the OCR-outputted document text passages could be created at some remote location and brought into the OCR output text triage system 200 through a transportable memory interface or from some other device connected to the network 110 which generates OCR-outputted document text passages.

[0029]

Fig. 2 illustrates a functional block diagram of one exemplary embodiment of the

OCR output text triage system 200. The OCR output text triage system 200 connects to the network 110 via the link 214. The link 214 can be any known or later developed device or system for connecting the OCR output text triage system 200 to the network 110, including a connection over public switched telephone network, a direct cable connection, a connection over a wide area network, a local area network or a storage area network, a connection over an intranet or an extranet, a connection over the Internet, or a connection over any other distributed processing network or system. In general, the link 214 can be any known or later-developed connection system or structure usable to connect the OCR output text triage system 200 to the network 110.

[0030] As shown in Fig. 2, the OCR output text triage system 200 includes one or more display devices 280 usable to display information to the user, and one or more user input devices 290 usable to allow the user or users to input data into the OCR output text triage system 200. The one or more display devices 280 and the one or more input devices 290 are connected to the OCR output text triage system 200 through an input/output interface 210 via one or more communication links 282 and 292, respectively, which are generally similar to the link 214 above.

[0031] In various exemplary embodiments, the OCR output text triage system 200 includes one or more of a controller 220, a memory 230, a trained off-line triage model 232, a text passage error threshold operating point model 234, an OCR-output text passage character attribute determination circuit or routine 240, an OCR-output text passage character accuracy determination circuit or routine 250, an OCR-output text passage accuracy determination circuit or routine 260, and an OCR-output text passage triage circuit or routine 270, all of which are interconnected over one or more data and/or control buses and/or application programming interfaces 295. In various exemplary embodiments, the trained off-line triage model 232 and the text passage error threshold operating point model 234 are stored in memory 230 of the OCR output text triage system 200.

[0032] The controller 220 controls the operation of the other components of the OCR output text triage system 200. The controller 220 also controls the flow of data between components of the OCR output text triage system 200 as needed. The

memory 230 can store information coming into or going out of the OCR output text triage system 200, may store any necessary programs and/or data implementing the functions of the OCR output text triage system 200, and/or may store data and/or OCR output text triage information at various stages of processing.

[0033] The memory 230 can be implemented using any appropriate combination of alterable, volatile or non-volatile memory or non-alterable, or fixed, memory. The alterable memory, whether volatile or non-volatile, can be implemented using any one or more of static or dynamic RAM, a floppy disk and disk drive, a writable or re-rewriteable optical disk and disk drive, a hard drive, flash memory or the like. Similarly, the non-alterable or fixed memory can be implemented using any one or more of ROM, PROM, EPROM, EEPROM, an optical ROM disk, such as a CD-ROM or DVD-ROM disk, and disk drive or the like.

[0034] In various exemplary embodiments, the OCR output text triage system 200 includes the trained off-line triage model 232 which the OCR output text triage system 200 uses to process a set of OCR-output document text passages using the various circuits or routines 240, 250, 260 and/or 270 to automatically triage an OCR-outputted document text passage. The trained off-line triage model 232 is trained on a large sample of OCR-output characters contained within a large set of OCR-output text passages that had been evaluated, for example, manually proofed. The trained off-line triage model 232 is discussed in detail below.

[0035] In various exemplary embodiments, the OCR output text triage system 200 includes a text passage error threshold operating point model 234 which the OCR output text triage system 200 uses to process a set of OCR-output document text passages using the various circuits or routines 240, 250, 260 and/or 270 to automatically triage an OCR-outputted document text passage. The text passage error threshold operating point model 234 is used to select a threshold operating point that will, with high confidence, satisfy customer-specified quality requirements while minimizing the labor needed to process document text passages that are not triaged. The text passage error threshold operating point model 234 is discussed in detail below.

[0036] The OCR-output text passage character attribute determination circuit or routine

240 is activated by the controller 220 evaluate OCR-output characters, shown as 170 in Fig. 10, contained in the OCR-output text passage. In various exemplary embodiments, a character may be expressed in several ways, such as, for example, a single utf8 multibyte representation of a single Unicode character which also includes all printable 7bit ASCII characters, as an "entity reference" such as "&omega", "&times", and "-", or the like.

[0037] In various exemplary embodiments, the OCR-output text passage character attribute determination circuit or routine 240 identifies and selects, as shown in Fig. 4, specific OCR-output text passage character attributes 160, such as, for example, the character class, the confidence descriptor class provided by the OCR system, the language of the text passage, the date of publication of the document, the typeface in which the text passage is printed, image-based features of the individual character image and/or surroundings, metadata attached to the document, that may be present in the OCR-output text passage 150.

[0038] In various exemplary embodiments, the OCR-output text character attribute determination circuit or routine 240 identifies and selects all or only a subset of the OCR-output text character attributes, such as the character class 161 and the confidence descriptor class 162, as shown in Fig. 4, from the large group of potential OCR-output text character attributes 160, such as the character class, the confidence descriptor class provided by the OCR system, the language of the text passage, the date of publication of the document, the typeface in which the text passage is printed, image-based features of the individual character image and/or surroundings, metadata attached to the document, that are available to use in triaging OCR-outputted text passages.

[0039]

The OCR-output text passage character accuracy determination circuit or routine 250 is activated by the controller 220 to determine, for each OCR-output character present in the OCR-output text passage, how accurately the OCR-output character is being interpreted by the OCR system. In various exemplary embodiments, the OCR-output text passage character accuracy determination circuit or routine 250 determines a character interpretation error value, such as, for example, a probability of error per character, $p_{(\text{character error})}$ using models contained in the trained off-

line triage model 232. In various exemplary embodiments, the determination of the character interpretation error value, such as, for example, the probability of error per character, $p_{(\text{character error})}$, includes at least a determination of at least one OCR-output character attribute being erroneously interpreted by the OCR system, such as, for example a probability, $p_{(\text{character attribute error})}$, of at least one OCR-output character attribute being erroneously interpreted by the OCR system

[0040] The OCR-output text passage character accuracy determination circuit or routine 250 processes each OCR character, with its one or more character attributes selected, through the trained off-line triage model 232 to determine the character interpretation error value, such as, for example, the probability of error per character, $p_{(\text{character error})}$. In various exemplary embodiments, the OCR-output text passage character accuracy determination circuit or routine 250 determines the character interpretation error value, such as, for example, the probability of error per character, $p_{(\text{character error})}$, using one or more statistical algorithms or methods. In one exemplary embodiment, the OCR-output text passage character accuracy determination circuit or routine 250 determines the character interpretation error value, such as, for example, the probability of error per character, $p_{(\text{character error})}$, using one or more latent conditional independence (LCI) statistical models included in the trained off-line triage model 232.

[0041] It should be noted that other known or later-developed statistical processes may be employed to process each OCR character with its attribute(s), including, for example, using one or more of language models such as character or word n-grams, Bayesian networks or other complex models of statistical dependence, and models of OCR error patterns.

[0042]

The OCR-output text passage accuracy determination circuit or routine 260 is activated by the controller 220 to determine one or more OCR-output text passage-wise quality metrics or scores using various statistical algorithms or methods included in the trained off-line triage model 232. In various exemplary embodiments, the OCR-output text passage accuracy determination circuit or routine 260 determines one or more OCR-output text passage-wise quality metrics, such as, for example, a text passage error value represented as a probability, $p_{(\text{text passage error})}$, that the

entire OCR-output text passage is erroneously interpreted by the OCR system, an OCR-output text passage error rate, $R_{(\text{text passage error rate})}$, or the like. In various exemplary embodiments, the OCR-output passage accuracy determination circuit or routine 260 determines the OCR-output text passage error rate, $R_{(\text{text passage error rate})}$, by combining, for example summing up, the error probabilities determined for each character in a text passage, $\Sigma p_{(\text{character error})}$, and then dividing the sum of error probability values, $\Sigma p_{(\text{character error})}$, by the number, N, of characters contained in the text passage.

[0043] The OCR-output text passage triage circuit or routine 270 is activated by the controller 220 to automatically perform one or more text passage triage decisions. In various exemplary embodiments, the OCR-output text passage triage circuit or routine 270 automatically performs one or more text passage triage decisions by comparing the one or more determined OCR-output text passage-wise quality metrics, such as, for example, the OCR-output text passage error rate, $R_{(\text{text passage error rate})}$, against a predetermined OCR-output text passage threshold error rate, $R_{(\text{text passage threshold error rate})}$, included in the text passage error threshold operating point model 234.

[0044] Based on the results of these error rate comparisons, the OCR-output text passage triage circuit or routine 270 also automatically determines the best post-OCR processing step required for that particular text passage. Post-OCR processing steps may include, for example, sending the OCR text passage directly to the end user without any post-OCR rekeying or correction, sending the OCR text passage through a post-OCR inspection and correction stage, sending the original text passage image to be completely keyed in manually, or a combination thereof. The determination of what post-OCR processing step may be taken is based at least on predetermined threshold error rate values, for example, as provided by a client.

[0045] Fig. 5 shows in greater detail one exemplary embodiment of the trained off-line triage model 232. Fig. 7 is a high-level schematic representation of one exemplary embodiment of the implementation stages or blocks of the method for triage of OCR-output text passages. As shown in Fig. 7, block 620 is an exemplary embodiment of a method of developing a trained off-line triage model 232.

[0046] As shown in Fig. 5, in one exemplary embodiment, the trained off-line triage model 232 includes an OCR-output text passage character attribute evaluation model 2322, an OCR-output text passage character labeling model 2324, a predictive model 2326 usable to determine probability distributions for OCR-output characters being correctly or erroneously interpreted by the OCR system, and a validation model 2328 usable to test, characterize and/or improve the developed predictive model 2326.

[0047] In various exemplary embodiments, the OCR-output text passage character attribute evaluation model 2322 determines, for example identifies and selects, at least one exemplary character attribute in an OCR-output text passage from a plurality of character attributes. As shown in Fig. 4, exemplary character attributes 160 include character class, confidence descriptor class, for example, confidence score provided by the OCR system, language of the text passage, date of publication of the document, typeface in which the text passage is printed, image-based features of the individual character image and/or surrounding, and metadata attached to the document. In an exemplary embodiment, as shown in Fig. 7, the OCR-output text passage character attribute evaluation model 2322 selects for evaluation two character attributes, a character class 161 and a confidence descriptor class or confidence score 162, from the plurality of character attributes 160. It will be noted that OCR systems generally attach confidence scores to each character interpretation. These confidence scores are, typically, integers within a narrow range of values. In an exemplary embodiment, the confidence scores assigned to OCR-output characters by an OCR system, for example ScanSoft"s TextBridge® OCR system, are in a range of 0 to 255.

[0048] In various exemplary embodiments, the OCR-output text passage character labeling model 2324 is used to label each OCR character in a training text passage data set as correct or incorrect using one or more algorithms or methods. In various exemplary embodiments, the OCR-output text passage character labeling model 2324 evaluates, at the character level, a training data set, shown as 622 in Fig. 7, consisting of OCR-output text. The OCR-output text passage character labeling model 2324 then compares the training data set 622 against a corresponding "ground-truth" or correct text. It will be noted that because of limitations associated with the OCR process and technology, the OCR-output text in the training set is generally errorful,

while the corresponding ground-truth text is generally highly accurate following a manual proofing step. Ground-truth text is commonly available in large quantities in such service bureaus, since it is the routine finished product of the process.

[0049] In one embodiment, the OCR-output text passage character labeling model 2324 is used to label each OCR character as "correct" or "error", shown as 623 in Fig. 7, using a string matching algorithm model, such as the model outlined in "The string to string correction problem," R. A. Wagner et al., Journal of Association of Computing Machinery, 21:168-178, 1974, incorporated herein by reference in its entirety.

[0050] It should be noted that other known or later-developed algorithms or methods may be employed to evaluate, label and/or compare each OCR character as "correct" or "error", including, for example, using a dynamic programming process, such as Unix "diff" programming process, or a more sophisticated process that takes into account models of OCR error patterns and human error patterns in ground-truth generation.

[0051] In various exemplary embodiments, once each OCR character in the training data set is automatically labeled as "correct" or "error," the model 2326 is used to produce statistical models of the relationship between the OCR accuracy and the attributes of a particular OCR-output character. In various exemplary embodiments, the model 2326 is used to produce predictive statistical models of class conditional character attributes, shown as 625 in Fig. 7, based on one or more latent conditional independence (LCI) statistical techniques or methods. In various exemplary embodiments, the model 2326 produces LCI-based statistical models 625 of the OCR-output character class and the OCR-output character confidence class or score, as best fitted by the data in the training data set, shown as 622 in Fig. 7, in terms of maximum likelihood sense, as described below.

[0052] In various exemplary embodiments, for each passage text character 170 having character attributes 160, the OCR system generates a string of pairs (m, n), shown as 163 in Fig. 7, where "m" 161 is a character attribute, and "n" 162 is a character confidence descriptor or confidence class. Each pair 163 is determined as either being "correct", for example the character attribute "m" 161 matches the ground-truth, or being "error", for example the character attribute "m" 161 does not match the ground-truth. Given the above observation,

the a posteriori probability of error p(error|m,n) can be determined using Equation 1 bel

$$p(\text{error}|m,n) = \frac{p(m,n\,|\,\text{error})p(\text{error})}{p(m,n\,|\,\text{error})p(\text{error}) + p(m,n\,|\,\text{correct})p(\text{correct})} \qquad (1)$$

ow:

[0053] In various exemplary embodiments, for each character, each of the two

conditional joint distributions over (m, n), conditioned on "correct" and "error," respectively, are modeled as latent conditionally independent (LCI) with K factors, where K is chosen experimentally to avoid over-fitting. The LCI models allow for different groups of character attributes, such as, for example, lower case, numeric and punctuation, to have different distributions of confidence classes or scores.

[0054] In various exemplary embodiments, at the OCR-output character level, at least one LCI-based model that best fits the data in the training data set in the maximum likelihood sense is determined for each of the two categories, "correct" and "error." In various exemplary embodiments, the at least one LCI-based model, shown as 625 in Fig. 7, is determined using various statistical algorithms or methods. In an exemplary embodiment, the at least one LCI-based model is determined using an Expectation Maximization (EM) algorithm model, such as the model outlined in "Maximum likelihood from incomplete data via the EM algorithm," A.P. Dempster et al., Journal of the Royal Statistical Society, 39(1), pp. 1-38, 1977, incorporated herein by reference in its entirety.

[0055] In various exemplary embodiments, the statistics for determining the multinomial LCI-based triage models may include, for example the number of occurrences of each (m, n) pair for each "correct" or "error" category. In one exemplary embodiment, the predictive statistical models of class conditional character attributes for triage of text passages, shown as 625 in Fig. 7, may be determined by performing a fixed number of iterations of the EM algorithm model without directly measuring convergence criteria.

[0056]

In various exemplary embodiments, the predictive model validation model 2328 is used to empirically validate the predictive statistical models, shown as 625 in Fig. 7, of class conditional character attributes against a validation data set. The predictive model validation model 2328 processes a validation set of text passage data, which were previously labeled with respect to their character accuracy, through the predictive models 625 to generate one or more operating characteristics. In one exemplary embodiment, the predictive model validation model 2328 processes a validation set of text passage data through the predictive models 625 to generate one or more operating characteristics that empirically quantify the trade off between the

"triage rate", which affects operating costs, and the "false hit rate", which affects the probability that a customer quality standard will be met.

[0057] In various exemplary embodiments, generating one or more operating characteristics that empirically quantify the trade off between the triage rate and the false hit rate facilitates the determination of a choice of a threshold operating point, shown as 662 in Fig. 7. The threshold operating point 662 may be used during text passage triage to maximize the number of text passages triaged while maintaining a reasonably low risk that too many text passages will exceed a predetermined error target. It will be noted that in an exemplary embodiment, the predetermined error target is set by a service bureau customer, while an acceptable risk level value is chosen by the service bureau managers.

[0058] Figure 6 is a flowchart outlining one exemplary embodiment of a method for creating or "training" an off-line triage model as a function of the OCR-output character attributes according to this invention. As shown in Fig. 6, the method begins in step S100, and continues to step S110, where one or more character attributes contained in an OCR-output text passage are selected from a plurality of character attributes. In various exemplary embodiments, the character attributes selected include the character class and the confidence descriptor class provided by the OCR system.

[0059] Then, in step S120, each OCR character in a training text passage data set of the off-line triage model is labeled as correct or incorrect using one or more algorithms or methods. In various exemplary embodiments, the training data set, consisting of OCR-output text, is evaluated and compared, at the character level, to corresponding "ground-truth" or correct text. Each OCR character is then labeled as "correct" or "error" using a string matching algorithm model, such as the model outlined in "The string to string correction problem," R. A. Wagner et al., Journal of Association of Computing Machinery, 21:168-178, 1974, incorporated herein by reference in its entirety. It should be noted that other known or later-developed algorithms or methods may be employed to evaluate, label and/or compare each OCR character as "correct" or "error", including, for example, using a dynamic programming process, such as Unix "diff" programming process, or a more sophisticated process that takes

into account models of OCR error patterns and human error patterns in ground-truth generation.

[0060] Then, in step S130, once each OCR character is labeled and encoded as "correct" or "error," statistical models representing a relationship between the OCR accuracy and the attributes of a particular OCR-output character are determined. In various exemplary embodiments, predictive statistical models of class conditional character attributes are determined using one or more latent conditional independence (LCI) statistical techniques or methods. In an exemplary embodiment, predictive LCI-based models are determined using an Expectation Maximization (EM) algorithm model, such as the model outlined in "Maximum likelihood from incomplete data via the EM algorithm," A.P. Dempster et al., Journal of the Royal Statistical Society, 39(1), pp. 1-38, 1977, incorporated herein by reference in its entirety. In an exemplary embodiment, predictive models usable to estimate probability distributions for an OCR-output character being correctly or erroneously interpreted by an OCR system are determined.

[0061] Next, in step S140, the developed LCI-based models usable to triage text passage characters, and thus text passages, are validated, such as for example, tested, characterized and/or improved. Operation then continues to step S150, where the operation of the training method stops.

[0062] In various exemplary embodiment, the validation of the LCI-based models may include processing a set of text passages that have been labeled with their "true quality' scores, as computed using a customer quality standard, through the developed predictive models. It will be noted that the text passages in the validation set are different from the text passages used in "training" the predictive models.

[0063] In various exemplary embodiments, the result of validation is an "operating curve" describing how well the triage model works for each setting of a "triage threshold" value. The performance of the triage model is measured in two ways: (1) the "triage rate," which affects operating costs, and (2) the "false hit rate," which affects the probability that the customer quality standard will be met. In an exemplary embodiment, an operation manager may examine the operating curve to select a threshold that will, with high confidence, meet customer accuracy requirements while

minimizing operating costs.

[0064] Fig. 7 is a high-level schematic representation of one exemplary embodiment of the implementation stages or blocks of the method for triage of OCR-output text passages according to this invention. Fig. 8 is a flowchart outlining one exemplary embodiment of a method for triage of an OCR-output text passage according to this invention.

[0065] As shown in Fig. 8, the method begins in step S200, and continues to step S210, where each OCR-output text passage is evaluated using the trained off-line triage model. In various exemplary embodiments, the OCR-output text passage is evaluated by determining at least one OCR-output character attribute for each OCR-output character in the OCR-output text passage. In one exemplary embodiment, determining at least one OCR-output character attribute for each OCR-output character includes selecting at least one OCR-output character attribute from a plurality of OCR-output character attributes. As described above, in one exemplary embodiment, the plurality of OCR-output character attributes may include at least some of character classes, confidence descriptor classes, languages of text passage, text passage publication date, typefaces in which text passages are printed, image-based features of an individual character image and metadata attached to text passages.

[0066] Then, in step S220, for each evaluated OCR-output text passage, a text passage error rate is determined using the trained off-line triage model and the determined OCR-output character attribute(s) in the OCR-output text passage. Next, in step S230, the error rate determined for each OCR-output text passage is compared with a predetermined OCR-output text passage threshold error rate. Based on the results of this comparison, an OCR-output text passage triage decision is made.

[0067] In various exemplary embodiments, the OCR-output text passage triage decision may include at least one of sending the OCR-output text passage directly to an end user without post-OCR rekeying or correction, shown as step S240 in Fig. 8, sending the OCR-output text passage through a post-OCR inspection and correction stage, shown as step S250, sending the original text passage image to be completely keyed in manually, shown as Step S255, or a combination thereof. In the exemplary

embodiments where the OCR-output text passage is provided to the user without post-OCR correction S240 or the OCR-output text passage is sent to be manually rekeyed S255, operation continues to step S300, where the operation of the OCR-output triage method stops.

[0068] In various exemplary embodiments, for the triage decision of sending the OCR-output text passage through a post-OCR inspection and correction stage, step S250, the method continues in step S260 where at least a portion of the OCR-output text passage is sent to be corrected after being identified as requiring correction by the OCR-output text triage system. Next, at step S270, the corrected OCR-output text passage is provided again to the triage system for evaluation. A corrected OCR-output text passage error rate is next determined and compared to the predetermined OCR-output text passage threshold error rate. If the error rate of the corrected OCR-output text passage is less than predetermined threshold error rate, the corrected OCR-output text passage is provided to the end user, as shown in step S290 in Fig. 8. Operation then continues to step S300, where the operation of the OCR-output triage method stops.

[0069] Fig. 9 is a flowchart outlining one exemplary embodiment of step S220, determining a text passage error rate for each OCR-output text passage evaluated using a trained off-line triage model based on the determined at least one OCR-output character attribute. As shown in Fig. 9, the method begins in step S222, and continues to step S224, where each OCR-output character, with its selected attributes, is provided to the trained off-line triage model.

[0070]

Next, in step S224, for each OCR-output character present in the OCR-output text passage, a determination is made of how accurately each OCR-output character is being interpreted by the OCR system. In various exemplary embodiments, this determination is performed by determining a character interpretation error value, such as, for example, a probability of error per character, $p_{(\text{character error})}$ using models contained in the trained off-line triage model. In various exemplary embodiments, the determination of the character interpretation error value, such as, for example, the probability of error per character, $p_{(\text{character error})}$, includes at least a determination of at least one OCR-output character attribute being erroneously

interpreted by the OCR system, such as, for example a probability, $p_{(\text{character attribute error})}$, of at least one OCR-output character attribute being erroneously interpreted by the OCR system

[0071] In various exemplary embodiments, the character interpretation error value, such as, for example, the probability of error per character, $p_{(\text{character error})}$, is determined using one or more statistical algorithms or methods in the trained off-line triage model. In one exemplary embodiment, the character interpretation error value, such as, for example, the probability of error per character, $p_{(\text{character error})}$, is determined using one or more latent conditional independence (LCI) statistical models included in the trained off-line triage model.

[0072] It should be noted that other known or later-developed statistical processes may be employed to process each OCR character with its attribute(s), including, for example, using one or more of language models such as character or word attribute n-grams, Bayesian networks or other complex models of statistical dependence, and models of OCR error patterns.

[0073] Then, in step S226, one or more OCR-output text passage-wise quality metrics or scores are determined using various statistical algorithms or methods included in the trained off-line triage model. In various exemplary embodiments, one or more OCR-output text passage-wise quality metrics are determined, such as, for example, a text passage error value represented as a probability, $p_{(\text{text passage error})}$, that the entire OCR-output text passage is erroneously interpreted by the OCR system, an OCR-output text passage error rate, $R_{(\text{text passage error rate})}$, or the like.

[0074] In various exemplary embodiments, the text passage error value, for example the probability $p_{(\text{text passage error})}$, is determined by combining the error probability values determined for each character in the text passage, $\Sigma\, p_{(\text{character error})}$. In various exemplary embodiments, the OCR-output text passage error rate, $R_{(\text{text passage error rate})}$, is determined by combining, for example summing up, the error probabilities determined for each character in a text passage, $\Sigma\, p_{(\text{character error})}$, and then dividing the sum of error probability values, $\Sigma\, p_{(\text{character error})}$, by the number, N, of characters contained in the text passage.

[0075] Operation then continues to step S229, where the operation of the method stops.

[0076] The method and systems according to this invention were implemented and their accuracy evaluated by the inventors. Specifically, 1413 scanned pages were processed with the TextBridge ® OCR system at the inventors' facility. The text on the pages was manually keyed in to provide ground-truth data. The OCR output, such as, for example, character labels and confidence scores, for each page was then aligned to ground-truth text using dynamic programming, for example Unix "diff", to obtain "error" or "correct" labels for each character label-confidence score pair.

[0077] The 1413 pages were then randomly permuted and partitioned into two samples of 706 and 707 pages respectively. The first sample was used for training the LCI models for "error" and "correct" groups, each model having 5 factors, for example K = 5. The second sample was used for validation.

[0078]

Table 1 shows the operating characteristics, such as, for example, false-hit rate vs. triage rate on the validation set. The false hit and triage rates were computed for different values of the threshold average a posteriori probability of error and a target quality of ≤ 1.5% OCR error. INSER

[t1]

Table 1: Operating characteristics on validation data : LCI model based triaging method.

Operating threshold	Triage rate (% of total)	False hit rate rate%	Good and triaged	Bad but triaged ('false hits')	Good but manually corrected ('false alarms')	Bad and manually corrected	Total tested
0.002	0.00	0.00	0	0	469	238	707
0.004	3.25	0.00	23	0	446	238	707
0.006	11.60	0.00	82	0	387	238	707
0.008	22.21	0.00	157	0	312	238	707
0.010	35.08	0.85	242	6	227	232	707
0.011	40.88	1.27	280	9	189	229	707
0.012	47.38	2.55	317	18	152	220	707
0.014	55.73	4.95	359	35	110	203	707
0.016	67.33	8.35	417	59	52	179	707
0.018	74.26	12.16	439	86	30	152	707

T TABLE

[0079] Of the 707 pages in the validation set, OCR output on 469 pages, or about 66%, met the target quality. With an operating threshold of 0.011 on the average a

posteriori probability of error, 289 pages, or about 41%, were triaged and bypass manual intervention. Of these, 280 were among the 469 "good quality" pages. Nine of the triaged pages did not actually meet target quality, representing a false hit rate of 1.3%, as an example of customer-specified tolerance.

[0080] Thus, the triage method of this invention, applied to this set of page images, has been shown capable of bypassing manual correction for 41% of the document stream fully automatically and without compromising the quality goals established by the customer.

[0081] Although the systems and methods for triage of OCR-output text passages according to this invention have been described in the context of using only a two-dimensional model, where the dimensions are character class and confidence class or score, the systems and methods for triage of OCR-output text passages are not limited to such two-dimensional models. The LCI model training and classification algorithms included in the trained off-line triage model can cope with higher dimensions. Thus, knowledge of any other characteristic of documents, such as, for example, including typeface, language, content topic, and image quality, may be included in the determinations performed using the trained off-line triage model.

[0082] The systems and methods for triage of OCR-output text passages according to this invention may be employed to automatically perform triage of OCR-output text passages, including for example, performing various "on-line" and "off-line" triage operations and decisions. In one exemplary embodiment of an "on-line" triage operation and decision, as the manual correction operator works down a page, overall page quality is continually re-estimated, so that the operator can be instructed to stop when the per-page accuracy target has, with sufficiently high statistical confidence, been reached.

[0083] In another exemplary embodiment, the systems and methods for triage of OCR-output text passages according to this invention may be employed to triage shorter passages, for example individual words, such that manual correction is directed to the most urgent corrections first. This exemplary embodiment and the immediately above-described embodiment are complementary and may be combined to enhance the capabilities of the systems and methods for triage of OCR-output text passages

according to this invention.

[0084] Although the invention has been described in detail, it will be apparent to those skilled in the art that various modifications may be made without departing from the scope of the invention.